2 Diagnosis

1. INTRODUCTION

The clinical decision-making process is based on probability. Based on certain clinical information such as risk factors, family history, and findings on physical examination, the clinician obtains some estimate of the probability of disease (the pretest probability). Diagnostic tests are performed in order to improve the estimate of this probability. If the test is negative, the probability of disease should fall and if the test is positive, then the probability of the disease should rise. Useful diagnostic tests will produce marked shifts in the probability of disease based on whether the results of the test are positive or negative.

In this chapter, we will explore the science of diagnostic testing from two perspectives. We shall first consider the methodological principles that underpin the design and conduct of studies of diagnostic tests. We shall then explore the epidemiological principles that guide the use and interpretation of diagnostic test results.

2. BASIC PRINCIPLES

Studies of the accuracy of diagnostic tests are generally of poor quality *(1)* and the ability to read and critically appraise the literature on diagnostic testing therefore requires some understanding of how to assess the quality of these studies. In evaluating the quality of a study of a diagnostic test, it is necessary to consider the process whereby participants were selected, the nature of the gold or reference standard that was used to determine the presence or absence of disease, whether the test under investigation (the index test) formed part of the reference standard, and whether the investigators who performed the index test were blind to the results of the reference standard. These issues have to do with the external validity of the study. That is to say, how easily can the results of this study be generalized to populations of patients encountered in clinical practice?

A diagnostic test is useful only insofar as it distinguishes between conditions that might otherwise be confused. A study that evaluates the ability of a test to distinguish two unrelated disorders or to distinguish individuals severely affected by a disease from healthy controls does not mimic the conditions encountered in clinical practice and so provides relatively little information about the real diagnostic value of the test. It is

From: *Neuromuscular Disease: Evidence and Analysis in Clinical Neurology*
By: M. Benatar © Humana Press Inc., Totowa, NJ

remarkable how often this basic principle is ignored in studies of diagnostic tests. This is a theme that will recur throughout this book. Furthermore, in reporting the results of a study of a diagnostic test, the authors should specify the setting in which subjects were recruited and whether study subjects represent a consecutive series or whether subjects were selected for inclusion on the basis of having received the index test or reference standard. Each of these factors are relevant to the external validity of the study.

In order to evaluate the performance of a diagnostic test it is also necessary to know something about the standard against which the results of the test were compared. In other words, it is necessary to know the criteria that were used as the gold or reference standard to determine whether the outcome of interest (e.g., disease) is present or absent. Reliable and robust gold standards do exist, but frequently they are not available in routine clinical practice. Pathological evidence for the presence of a disease might be considered the gold standard for certain diseases, but biopsy or autopsy material may not always be available, which makes it difficult to rely on pathology as the gold standard for a diagnostic test. As a compromise, investigators frequently rely on some set of clinical criteria in order to evaluate the accuracy of a diagnostic test. In doing so, it is crucial to avoid incorporating the results of the test of interest into the criteria that will be used as the surrogate gold standard (an error referred to as "incorporation bias"). Although this recommendation is intuitive based on common sense, it is ignored with surprising frequency (as shall be illustrated repeatedly throughout the course of this book).

Blinding of the investigator to the results of the reference standard is also of fundamental importance to the validity of a study of a diagnostic test. Knowledge of the results of the reference standard may influence (consciously or subconsciously) the interpretation of the index test and thus introduce significant bias.

Various methods for evaluating the quality of diagnostic studies have been reported *(2,3)*. Most prominent among these is the Standards for Reporting of Diagnostic Accuracy (STARD) initiative, which represents an attempt to improve the quality of studies that are designed to investigate the accuracy of diagnostic studies. Among the methodological issues highlighted by the STARD document are the need for a clear definition of the population under study with explicit inclusion and exclusion criteria, the need for clear articulation of the gold (reference) standard against which the diagnostic test is being compared, the necessity of blinding of the examiner to the results of the gold standard, and the need for a clear discussion of how indeterminate results were handled. The STARD recommendations have received broad acceptance and will likely play an important role in the definition of a new standard by which studies of diagnostic tests will be judged.

3. SENSITIVITY AND SPECIFICITY

It is extremely unlikely that a positive test will always imply the presence of disease and that a negative test will always indicate the absence of disease. Because every test is likely to be fallible, there are four possible outcomes for any test in which the result is reported only as either positive or negative (i.e., no indeterminate results). These possible results are illustrated in the following table.

	"Truth"	
Test result	*Disease*	*No disease*
Positive	a True positive (TP)	b False positive (FP)
Negative	c False negative (FN)	d True negative (TN)

Those with the disease for whom the test is positive are labeled as true positive (a).

Those without the disease for whom the test is positive are labeled as false positive (b).

Those with the disease for whom the test is negative are labeled as false negative (c).

Those without the disease for whom the test is negative are labeled as true negative (d).

The sensitivity of a diagnostic test is defined as the proportion of people with the disease who test positive. That is say, sensitivity = $\frac{a}{a+c} = \frac{TP}{TP+FN}$.

For a test with high sensitivity, most individuals with the disease will test positive. Because there are few false negatives, a negative test result provides good evidence that the disease is not present. Tests with high sensitivity are typically most useful for ruling out a particular disease. This is the sort of test that would be most useful as a screening test, in that a negative result reliably excludes disease. Subjects with positive results, however, will require more detailed testing to determine whether the disease is truly present.

The specificity of a diagnostic test is defined as the proportion of people without the disease who test negative. That is say, specificity = $\frac{d}{b+d} = \frac{TN}{TN+FP}$.

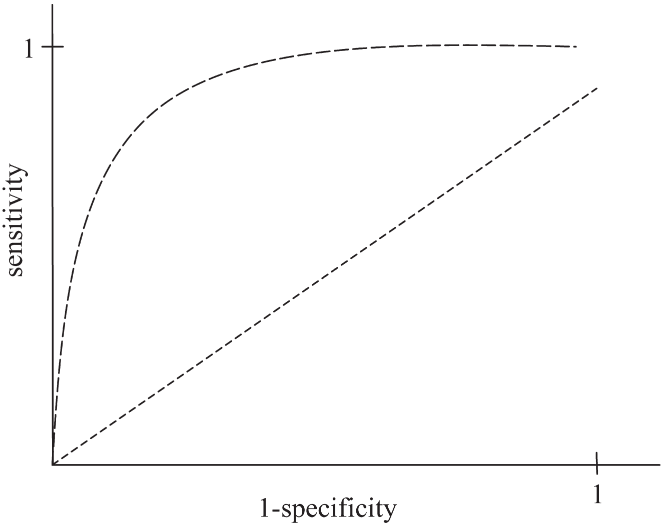
For a test with high specificity, most individuals without the disease will test negative. Because there are few false positives, a positive test result provides good evidence that the disease is present. Tests with high specificity are most useful for ruling in a particular disease.

Diagnostic accuracy is the term used to describe the combination of sensitivity and specificity. Accuracy can be estimated from the area under the receiver operating characteristic (ROC) curve (explained in "Cut-Points and ROC Curves").

Although sensitivity and specificity are properties of the diagnostic test, the performance of the test may vary among different populations *(4)*; hence the importance of evaluating the performance of the diagnostic test in the population of patients in which it will ultimately be used clinically. To illustrate how the performance of a test may vary from population to population, consider the example of the use of single-fiber electromyography (SFEMG) for the diagnosis of myasthenia gravis. Evaluation of the accuracy of SFEMG by comparing healthy subjects with those who have seropositive generalized myasthenia gravis may show that the test is both sensitive and specific. If the accuracy of SFEMG is instead evaluated using patients with mitochondrial disease or oculopharyngeal muscular dystrophy as controls and myasthenics with purely ocular disease as cases, both the sensitivity and specificity of the test might be expected to be reduced.

4. CUT-POINTS AND ROC CURVES

Sensitivity and specificity, therefore, are terms used to describe the performance of a test relative to some external gold standard. The sensitivity and specificity of a test depend, to a large extent, on the threshold or cut-point that is used to discriminate between a positive test result and a negative test result. In general, if the threshold for a positive (abnormal) test result is raised (i.e., making the test less likely to produce a positive result), then the sensitivity will fall and the specificity will rise. If, on the other hand, a low threshold for a positive test is used, then the test will be more likely to detect a greater proportion of subjects who have the disease (i.e., increased sensitivity), but this may come at the price of more frequent positive test results even among those without the disease (i.e., lower specificity). In general, good (i.e., useful) diagnostic tests combine high sensitivity with high specificity. The utility of a diagnostic test using different cut-points can be explored by plotting a ROC curve as shown in Graph 2.1., in which sensitivity on the *y*-axis is plotted against 1-specificity on the *x*-axis.



Graph 2.1

Tests in which improved sensitivity are accompanied by a fall in specificity of similar magnitude have no discriminative value. The ROC curve for such a test is illustrated by the straight 45° line. The diagnostic accuracy of such a test, which may be calculated from the area under the curve, is 50%. Such a test performs as well as the flip of a coin in determining whether disease is present or not.

Graph 2.1 shown above also illustrates the ROC curve for a test with almost perfect diagnostic accuracy. Specificity remains close to 100% at every test threshold until sensitivity also reaches almost 100%. This ROC curve ascends almost vertically along the *y*-axis from 0 to 1 and then extends almost horizontally along the *x*-axis from 0 to 1. The area under this curve is close to 100%, indicating that the test has an almost perfect discriminative value between disease and no disease. In reality, no test conforms to such specifications, but these hypothetical examples serve to illustrate the characteristics of a useful diagnostic test. The more toward the upper left-hand corner of the graph the ROC curve is located, the better the diagnostic accuracy of the test. The point of inflection of the ROC curve indicates the cut-point at which the combination of sensitivity and specificity is maximized.

5. PREDICTIVE VALUE

In contrast with sensitivity and specificity, which characterize the diagnostic accuracy of a test, the positive and negative predictive values are two estimates that directly address the probability of disease.

	"Truth"	
Test result	*Disease*	*No disease*
Positive	a True positive	b False positive
Negative	c False negative	d True negative

The positive predictive value represents the probability of disease being present if the test is positive. Using the hypothetical 2 × 2 table above, it is calculated as $\frac{a}{a+b} = \frac{TP}{TF+FP}$.

The negative predictive value similarly represents the probability of disease being absent if the test is negative. It is calculated as $\frac{d}{c+d} = \frac{TN}{TN+FN}$.

Use of the predictive values illustrates the point that the utility of a test depends on the population in which the test is being applied. The pretest probability of disease depends on the prevalence of disease within the specified population. The positive predictive value will be greater and the negative predictive value will be reduced in populations with high disease prevalence (i.e., high pretest probability).

6. LIKELIHOOD RATIOS

The likehood ratio (LR) is the ratio of the probability of a particular test result for a person with the disease divided by the probability of that same result for a person without the disease. The LR indicates by how much a given diagnostic test result will raise or lower the pretest probability of the disease in question.

The LR for a positive test (LR+) is defined as the probability of a positive test result for a person with the disease divided by the probability of a positive test result for a person without the disease. LR+ may be calculated as $\frac{sensitivity}{1 - specificity}$. Similarly, the LR for a negative test (LR–) is defined as the probability of a negative test result for a person with the disease divided by the probability of a negative test result for a person without the disease. LR– may be calculated as $\frac{1 - sensitivity}{specificity}$.

Test results may, for example, be categorized as indicating high probability of disease, moderate probability, low probability, or no probability of disease. Within each level of the test result it is possible to calculate the LR and to use this ratio to estimate the posttest probability of disease.

LRs ratios greater than 10 and less than 0.1 generate large and often definitive changes from pretest to posttest probability, LRs between 5 and 10 and between 0.1 and 0.2 lead to moderate changes in pretest to posttest probability, and LRs between 2 and 5 and between 0.2 and 0.5 result in small changes in probability. LRs between 1 and 2 and between 0.5-1 rarely alter pretest probability *(5)*.

The utility of the LR is that it can be used to estimate the posttest probability of disease. The strategy for doing so is illustrated as follows:

1. Determine the pretest probability of disease (this is typically based on an assessment of risk factors, family history, personal history, and physical examination).
2. Convert pretest probability to pretest odds (divide by 1 – pretest probability).
3. Multiple pretest odds by the LR to yield the posttest odds.
4. Convert the posttest odds to the posttest probability (divide by 1 + posttest odds).

7. CONCLUSION

The principles that should guide the conduct of a study of a diagnostic test should be born in mind when reading and evaluating the relevant literature. A study's failure to include an appropriate population, its susceptibility to incorporation bias, or the inadequacy of blinding should lead to caution in the interpretation of the results. Studies of diagnostic tests do not always frame their findings using the appropriate terminology (sensitivity, specificity, LR), but where possible, every effort should be made to tabulate data in such a way as to permit estimation of these parameters. So doing will permit the reader to think in terms of probability and to evaluate the utility of diagnostic tests in terms of a test's ability to move the posttest probability toward either very low or very high likelihood of disease.

REFERENCES

1. Reid M, Lachs M, Feinstein A. Use of methodological standards in diagnostic test research: getting better but still not good. JAMA 1995;274:645–651.
2. England J, Gronseth G, Franklin G, et al. Distal symmetric polyneuropathy: A definition for clinical research: Report of the American Academy of Neurology, the American Association of Electrodiagnostic Medicine, and the American Academy of Physical Medicine and Rehabilitation. Neurology 2005; 64:199–207.

3. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Br Med J 2003;326:41–44.
4. Jaeschke R, Guyatt G, Lijmer J. Diagnostic tests. In: (Guyatt G, Rennie D, eds.) User's Guide to the Medical Literature. A Manual for Evidence-Based Clinical Practice. American Medical Association Press, Chicago: 2002:121–140.
5. Jaeschke R, Guyatt G, Sackett D. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 1994;271:703–707.